iScience

CellPress
OPEN ACCESS

Perspective

Toward understanding the communication in sperm whales

Jacob Andreas,[1,16] Gašper Beguš,[2,16] Michael M. Bronstein,[3,4,5,16,*] Roee Diamant,[6,16] Denley Delaney,[7,16] Shane Gero,[8,9,16] Shafi Goldwasser,[10] David F. Gruber,[11,16] Sarah de Haas,[12,16] Peter Malkin,[12,16] Nikolay Pavlov,[16] Roger Payne,[16] Giovanni Petri,[13,16] Daniela Rus,[1,16] Pratyusha Sharma,[1,16] Dan Tchernov,[6,16] Pernille Tønnesen,[14,16] Antonio Torralba,[1,16] Daniel Vogt,[15,16] and Robert J. Wood[15,16]

SUMMARY

Machine learning has been advancing dramatically over the past decade. Most strides are human-based applications due to the availability of large-scale datasets; however, opportunities are ripe to apply this technology to more deeply understand non-human communication. We detail a scientific roadmap for advancing the understanding of communication of whales that can be built further upon as a template to decipher other forms of animal and non-human communication. Sperm whales, with their highly developed neuroanatomical features, cognitive abilities, social structures, and discrete click-based encoding make for an excellent model for advanced tools that can be applied to other animals in the future. We outline the key elements required for the collection and processing of massive datasets, detecting basic communication units and language-like higher-level structures, and validating models through interactive playback experiments. The technological capabilities developed by such an undertaking hold potential for cross-applications in broader communities investigating non-human communication and behavioral research.

INTRODUCTION

For centuries, humans have been fascinated by how animals communicate (Fögen, 2014). Animals use signals to communicate with conspecifics for a variety of purposes throughout their daily routines; yet it has been argued that their communication systems are not comparable, quantitatively or qualitatively, to human languages (Hauser et al., 2002). The latter derive their expressive power from a number of distinctive structural features, including displacement, productivity, reflexivity, and recursion. Whether known non-human communication systems exhibit similarly rich structure—either of the same kind as human languages, or completely new—remains an open question.

Understanding language-like communication systems requires answering three key technical questions: First, by analogy to the phonetics and phonology of human languages, what are the *articulatory and perceptual building blocks* that can be reliably produced and recognized? Second, by analogy to the morphology and syntax of human languages, what are the *composition rules* according to which articulatory primitives can be structurally combined? Third, by analogy to semantics in human languages, what are the interpretation rules that assign *meanings* to these building blocks? Finally, there may possibly be a *pragmatics* component, whereby meaning is additionally formed by *context* (Schlenker et al., 2016). While individual pieces of these questions have been asked about certain animal communication schemes, a general-purpose, automated, large-scale data-driven toolkit that can be applied to non-human communication is currently not available.

The recent success of machine learning (ML) methods in answering similar questions in human languages (Natural Language Processing or NLP) is related to the availability of large-scale datasets. The effort of creating a biological dataset in a format, level of detail, scale, and time span amenable to ML-based analysis is capital intensive and necessitates a multidisciplinary expertise to develop, deploy, and maintain specialized hardware to collect acoustic and behavioral signals, as well as software to process and analyze them, develop linguistic models that reveal the structure of animal communication and ground it in

[1]MIT CSAIL, Cambridge, MA, USA

[2]Department of Linguistics, University of California, Berkeley, CA, USA

[3]Department of Computer Science, University of Oxford, Oxford, UK

[4]IDSIA, University of Lugano, Lugano, Switzerland

[5]Twitter, London, UK

[6]Leon H. Charney School of Marine Sciences, University of Haifa, Haifa, Israel

[7]Exploration Technology Lab, National Geographic Society, Washington DC, USA

[8]Dominica Sperm Whale Project, Roseau, Commonwealth of Dominica

[9]Department of Biology, Carleton University, Ottawa, ON, Canada

[10]Simons Institute for the Theory of Computing, University of California, Berkeley, CA, USA

[11]Department of Natural Sciences, Baruch College and The Graduate Center, PhD Program in Biology, City University of New York, New York, NY, USA

[12]Google Research, Mountain View, CA USA

[13]ISI Foundation, Turin, Italy

[14]Marine Bioacoustics Lab, Zoophysiology, Department of Biology, Aarhus University, Aarhus, Denmark

[15]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

[16]Project CETI, New York, NY, USA

*Correspondence: michael.bronstein@cs.ox.ac.uk

https://doi.org/10.1016/j.isci.2022.104393

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

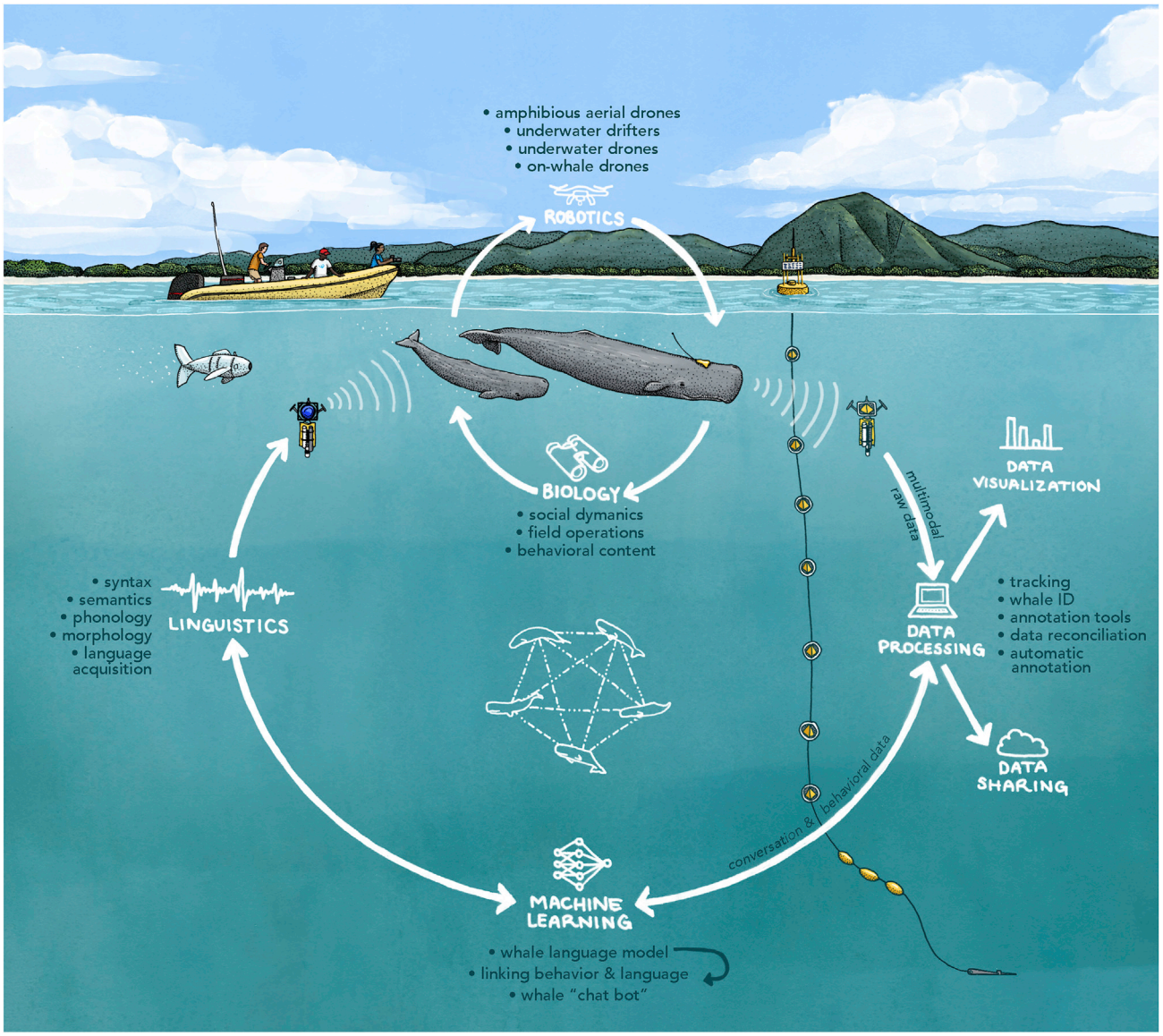

Figure 1. An approach to sperm whale communication that integrates biology, robotics, machine learning, and linguistics expertise, and comprise the following key steps

Record: collect large-scale longitudinal multimodal dataset of whale communication and behavioral data from a variety of sensors. Process: reconcile and process the multi-sensor data. Decode: using machine learning techniques, create a model of whale communication, characterize its structure, and link it to behavior. Encode & Playback: conduct interactive playback experiments and refine the whale language model. Illustration © 2021 Alex Boersma.

behavior, and finally perform playback experiments to attempt bidirectional communication for validation (Figure 1). Yet, the deployment of graphics processing unit's (GPU) is following a trajectory akin to Moore's Law (https://openai.com/blog/ai-and-compute) and, at the same time, the success of such an endeavor could potentially yield cross-applications and advancements in broader communities investigating non-human communication and animal behavioral research. One of the main drivers of progress making deep learning successful has been the availability of large (both labeled and unlabeled) datasets (and of architectures capable of taking advantage of such large data). To build a more complete picture and capture the full range of a species' behavior, collecting datasets containing measurements across a broad set of factors is essential. In turn, setting up infrastructure that allows for the collection of broad and sizable datasets would facilitate studies that allow the autonomous discovery of the meaning-carrying units of communication.

A dedicated interdisciplinary initiative toward a detailed understanding of animal communication could arguably be made with a number of species as its focus. Birds, primates, and marine mammals have all given insight into the capacity of animal communication. In some ways, the collective understanding of the capacity for and faculty of communication in non-humans has been built through experimentation and observation across a wide number of taxa (Fitch, 2005; Hauser et al., 2002). The findings on both the underlying neurobiological systems underpinning communicative capacity, and the complexity and diversity of the communication system itself often mirror our ability with which to work with a given species, or the existence of prominent long-term field research programs.

Animal communication researchers have conducted extensive studies of various species, including spiders (*e.g.* Elias et al., 2012; Hebets et al., 2013), pollinators (*e.g* Kulahci et al., 2008), rodents (*e.g* Ackers and Slobodchikoff, 1999; Slobodchikoff et al., 2009), birds (*e.g* Baker, 2001; Griesser et al., 2018), primates (*e.g.* Clarke et al., 2006; Jones and Van Cantfort, 2007; Leavens, 2007; Ouattara et al., 2009; Schlenker et al., 2016; Seyfarth et al., 1980), and cetaceans (*e.g* Janik, 2014; Janik and Sayigh, 2013), showing that animal communication involves diverse strategies, functions, and hierarchical components, and encompasses multiple modalities. Previous research efforts often focused on the mechanistic, computational, and structural aspects of animal communication systems. In human care, there have been several successful attempts of establishing a dialogue with birds (*e.g.*((Pepperberg, 1990)) and primates through a shared, trained, anthropocentric lexicon or various media such as iconographic keyboards (*e.g.* (Savage-Rumbaugh et al., 1985) or sign language (*e.g.* (Patterson, 1978)). However, due to the complexity of the environment and logistical challenges, such studies are often limited in sample size, continuity, and duration.

A comparatively long list of skills required for language learning in humans has been demonstrated among cetaceans (whales, dolphins, and porpoises), who share many social characteristics that are strikingly similar to our own. Whales and dolphins are among a few animals capable of vocal production learning (the ability to copy novel sounds as well as to vary those to produce individually distinctive repertoires) in addition to some birds, bats, pinnipeds, and elephants (Janik and Slater, 1997, 1998; Poole et al., 2005). Of those, only a few species, including parrots and dolphins appear to use arbitrary, learned signals to label objects or conspecifics in their communities in the wild (Balsby and Bradbury, 2009; Janik and Slater, 1998; King and Janik, 2013; Tyack and Sayigh, 1997; Wanker et al., 2005). Dolphins can use learned vocal labels to refer to and address each other when they meet at sea (King and Janik, 2013; Quick and Janik, 2012). This sort of vocal recognition system mediates highly dynamic societies among cetaceans, which involve social relationships lasting decades as well as regular interaction with strangers (Bruck, 2013; Connor, 2000; Gero et al., 2015, 2016a; Tyack, 1986).

At the same time, cetaceans provide a dramatic contrast in their ecology and environment compared to terrestrial animals (Steele, 1985). The logistical and technological difficulties related to the observation of marine life are one of the reasons why relatively little is known about many of the toothed whales (Odontocetes). For example, it was not until 1957 that it was even noted that sperm whales (*Physeter macrocephalus)* produce sound (Worthington and Schevill, 1957) and only in the 1970s came the first understanding that they use sound for communication (Watkins and Schevill, 1977). Among all odontocetes species, *P. macrocephalus* stands out as an "animal of extremes" (Weilgart et al., 1996; Whitehead, 2003). Sperm whales are the largest of the toothed whales, among the deepest divers, and have a circumglobal distribution (Whitehead, 2003). They can be both ocean nomads and small island specialists whose homes are both thousands of kilometers across and thousands of meters deep (Cantor et al., 2019). Sperm whales' immense nose, the origin of their biological name ("macrocephalus" translates as "large head"), houses the world's most powerful biological sonar system, which in turn is controlled by the world's largest brain, six times heavier than a human one (Goldbogen and Madsen, 2018; Marino, 1998; Møhl et al., 2003) and with large cerebral hemispheres and spindle neurons (Butti et al., 2009; Hof et al., 2005; Marino, 2004; Marino et al., 2011). These cerebral structures might be indicative of complex cognition and higher-level functions put by sperm whales to task in both their rich social lives and their complex communication system.

When comparing marine species with their terrestrial counterparts, it is important to emphasize the scale of the ocean across all dimensions. Many whales journey thousands of kilometers (*e.g.* (Stevick et al., 2011) and some are thought to live longer than a hundred years (Seim et al., 2014). Compared to their terrestrial counterparts, marine species also experience substantially greater environmental variation over periods of months or longer (Steele, 1985), and effectively live in a three-dimensional environment (Haskell et al., 2002;

Wosniack et al., 2017), creating a situation in which social learning is favored over individual learning or genetic determination of behavior. Together with the fact that many cetaceans live in stable social groups with prolonged parental care, the opportunities for cultural transmission of information and traditional behaviors are high with traits being passed consistently within social groups, but less often between them. As a result, several cetacean species exhibit high levels of behavioral variation between social groups, much of which is thought to be due to social learning. The marine environment renders chemical signals less effective, and whales rely on acoustics as their primary mode of communication. Most of their communication is thus likely to be captured by a single modality, while vision probably plays a significant role while in the photic zone.

The problem of constructing an inventory of phonetic, lexical, or grammatical units is much harder for whale communication compared to human languages, not only do we not know which acoustic features are meaningful, how they vary systematically, or which units are correlated with behavior, it is also not trivial to probe for and test and verify meaningful units in their communication. Unaided, human experimenters would have to evaluate an extremely large number of hypotheses by hand. Automatically discovered feature representations or coda boundaries can reduce the number of possibilities that human researchers have to consider, providing initial proposals for phoneme or phrase boundaries that can guide higher-level human analysis (Suzuki et al., 2006).

While multiple efforts in past decades to analyze non-human communication have brought a significant new understanding of various animal species and the structure and function of their signals, we still largely lack a functional understanding of non-human communication systems. Unlike human languages that are "pre-segmented, i.e. where basic units are already available, in non-human communication, this is not the case. Identifying such elements among animals has traditionally been done slowly using manual expert annotation and rests on anthropocentric assumptions, whereas ML offers a scalable approach. In retrospect, we can conclude that critical understanding was acquired slowly across long periods of time invested with specific communities of a limited number of species and a modestly sized amount of data for each. This is contrasted with the rapid growth of technologies that allow one to collect and process huge amounts of data. One such technology is ML, in particular, deep learning (Lecun et al., 2015) that has had a dramatic impact in natural language processing (NLP). Over the past decade, advances in machine learning have provided new powerful tools to manipulate language, making it now possible to construct unsupervised human language models capable of accurately capturing numerous aspects of phonetics and phonology, syntax, sentence structure, and semantics. Today's state-of-the-art NLP tools can segment low-level phonetic information into phonemes, morphemes, and words (Elsner et al., 2012), turn word sequences into grammars (Kim et al., 2019; Klein and Manning, 2004; Naseem et al., 2010), and ground words in visual perception and action (Andreas et al., 2016; Rohrbach et al., 2016; Tellex et al., 2011). These NLP tools can be transferred from natural language to non-human vocalizations in order to identify patterns that would be difficult to discover with a manual analysis. However, interpretable models of both human language and animal communication rely on formal approaches and theoretical insights (Berwick et al., 2011; Davies et al., 2021; Schlenker et al., 2016; Stokes et al., 2020). ML outputs are thus primarily a tool to constrain hypothesis space based to build formal and interpretable descriptions of the sperm whale communication. Using ML models for constraining hypothesis space has already been successfully applied in the fields of pure mathematics (Davies et al., 2021), drug discovery (Stokes et al., 2020), or protein folding (Jumper et al., 2021). Combining key concepts from machine learning and linguistic theory could thus substantially advance the study of non-human communication and, more broadly, bring a data-centric paradigm shift to the study of animal communication.

In this paper, we describe the current state of knowledge on sperm whale communication and outline the key ingredients of the collection and processing of massive bioacoustic data from sperm whales, detecting their basic communication units, language-like higher-level features, and discourse structure. We discuss experiments required to validate linguistic models and attribute meaning to communication units, and conclude with perspectives about the future progress in the field.

BACKGROUND

Sperm whales are born into tightly knit matrilineal families within which females (who are not always related) and their offspring make group decisions when traveling (Whitehead, 2016), finding food, and foraging together (Whitehead, 2003). Family members communally defend and raise their offspring, including

nursing each others' calves (Gero et al., 2009, 2013; Whitehead, 2003). Some families join up for hours to a few days to form "groups" with evidence of decade-long associations (Gero et al., 2013). On a higher level, sperm whales form clans of up to hundreds to tens of thousands of individual whales and exhibit diversity in movement patterns, habitat use, diving synchronization, foraging tactics, and diet; these differences appear to impact survival (Cantor and Whitehead, 2015; Marcoux et al., 2007; Whitehead and Rendell, 2004). Sperm whale clans coexist in overlapping ranges but remain socially segregated, despite not being genetically distinct communities (Rendell et al., 2012).

Acoustic communication of sperm whales

Despite its present-day use for communication, the sperm whales' remarkable bioacoustic system (see Figure 2A) evolved as a sensory device for echolocation allowing the whales to find prey and navigate in the darkness of the deep ocean (Goldbogen and Madsen, 2018; Tønnesen et al., 2020). Each short, highly directional, broadband echolocation click has a multi-pulse structure with an intense first pulse followed by a few additional pulses of decaying amplitude (see Figure 2B). The multi-pulsed click is the result of the reverberation of the initial pulse in the whale's *spermaceti organ* within its nose (Møhl et al., 2003; Zimmer et al., 2005).

Whale communication utilizes short (<2 s) bursts of clicks produced in stereotyped patterns that can be classified into recognizable types termed *codas* (Watkins and Schevill, 1977; Weilgart and Whitehead, 1997) (see Figure 2B). Distinct vocal sperm whale dialects have been documented in the Pacific, Indian, and Atlantic oceans (Amano et al., 2014; Amorim et al., 2020; Gero et al., 2016b; Huijser et al., 2020; Rendell and Whitehead, 2003). Each distinct socially learned clan dialect contains at least 20 different coda types. A typical coda is made up of 2–40 broadband omnidirectional clicks. Codas are produced most prolifically during longer periods of intense socialization near the surface when sperm whales are in close contact, at the onset of deep foraging dives, as well as during ascent when approaching the surface, but not when at depth foraging (Watwood et al., 2006; Whitehead, 2003). Recent insights into the coda repertoires used by individuals and groups of whales have suggested that specific codas encode varying levels of social recognition to mediate the animals' complex multilevel societies (Gero et al., 2016b). Codas appear to be rich in information about the caller's identity and there is some understanding of the diversity of coda types and the patterns of variation in their usage. Yet, the communicative function of particular codas themselves is still largely a mystery.

Codas are exchanged in duet-like sequences between two or more sperm whales. There is apparent turn-taking with whales responding within 2 s of each other, often overlapping and matching identical calls (Schulz et al., 2008). These exchanges occur across spatial scales ranging from meters to kilometers, suggesting that they function both between whales immediately together and those farther apart. Individuals within a family share a natal dialect of at least 10 coda types, despite there being some variation in individual production repertoires (Gero et al., 2016b; Schulz et al., 2011). Calves take at least two years to produce recognizable coda types and appear to "babble" in producing a larger number of call types prior to narrowing their usage to the types produced by their natal family (Gero et al., 2016b).

Machine learning for automatic annotation and representation learning

The time and capital investment as well as technical and logistical challenges connected to collecting high-quality field audio recordings and subsequently manually annotating and analyzing them have been a key factor to the relatively slow pace in the study of sperm whale communication. Given these challenges, the development of improved computational techniques for automatic processing, annotation, and analysis of information content and communicative intent of whale vocalizations is a crucial step for future progress in the field. Machine learning techniques used for the analysis of human language (speech recognition and natural language processing) provide great potential in addressing these challenges. Encouraging results in this direction were shown by (Bermant et al., 2019; Zhong et al., 2020), who used ML methods to automatically detect clicks in whale vocalization recordings, distinguish between echolocation and communication clicks, and classify codas into clans and individuals, achieving accuracy similar to previous highly time-consuming manual annotations and older generation statistical techniques. Recent advances in unsupervised learning trained on either spectral representations or raw waveforms potentially allow the use of ML in more complex tasks such as self-supervised acoustic unit discovery, which can provide the crucial first step in understanding communication systems without known meaningful units and without an easy way to elicit meaning. Self-supervised learning is not only appropriate for such tasks, but has been shown to learn more robust representations than supervised learning (e.g. in the visual domain (Goyal et al., 2022).

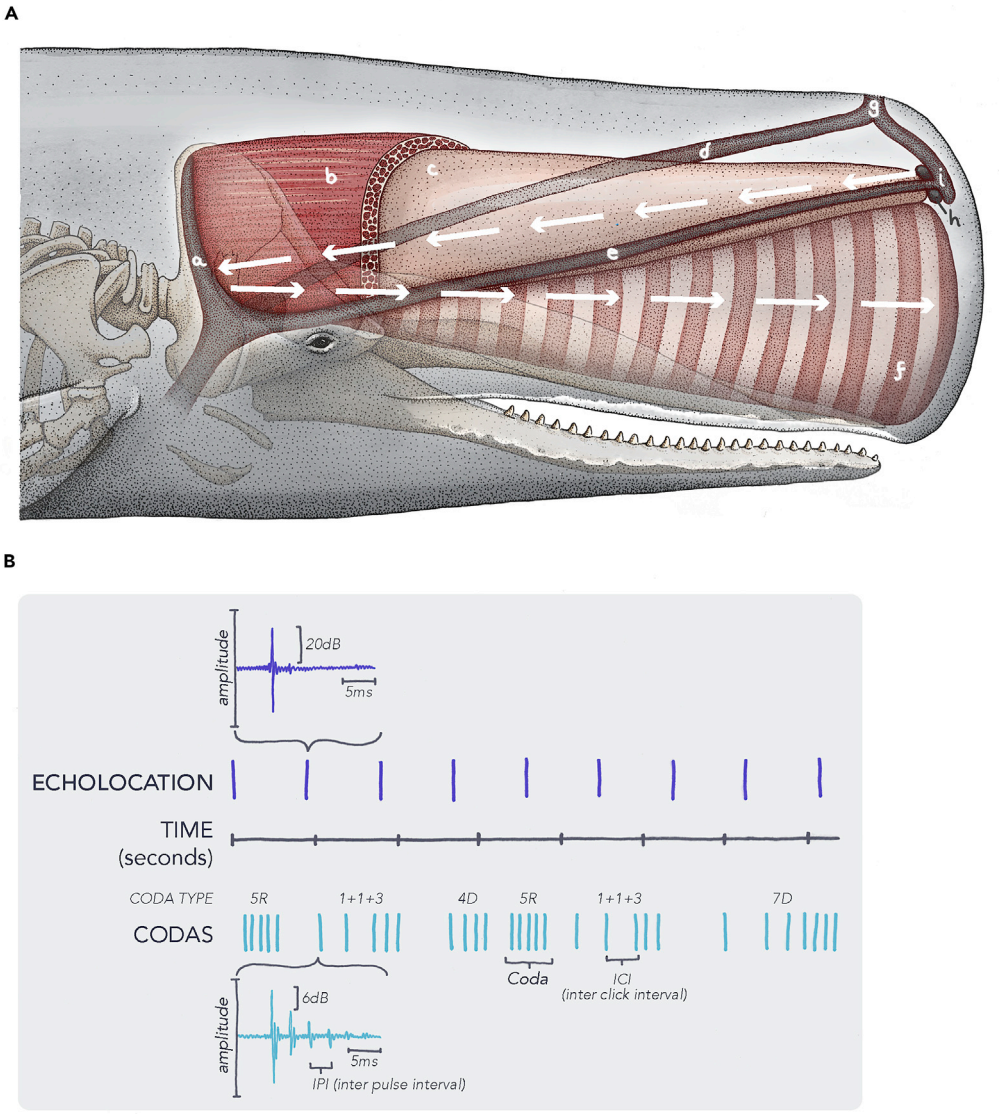

Figure 2. Sperm whale bioacoustic system

(A) Sperm whale head contains the spermaceti organ (c), a cavity filled with almost 2,000 L of wax-like liquid, and the junk compartment (f), comprising a series of wafer-like bodies believed to act as acoustic lenses. The spermaceti organ and junk act as two connected tubes, forming a bent, conical horn of about 10 m in length and 0.8 m aperture in large mature males. The sound emitted by the phonic lips (i) in the front of the head is focused by traveling through the bent horn, producing a flat wavefront at the exit surface.

(B) Typical temporal structure of sperm whale echolocation and coda clicks. Echolocation signals are produced with consistent inter-click intervals (of approximately 0.4 s) while coda clicks are arranged in stereotypical sequences called "codas" lasting less than 2 s. Codas are characterized by the different number of constituent clicks and the intervals between them (called inter-click intervals or ICIs). Codas are typically produced in multi-party exchanges that can last from about 10 s to over half an hour. Each click, in turn, presents itself as a sequence of equally spaced pulses, with inter-pulse interval (IPI) of an order of 3–4 ms in an adult female, which is the result of the sound reflecting within the spermaceti organ. Illustration © 2021 Alex Boersma.

Today's ML systems used in natural language processing applications are predominantly based on *deep representation learning*: input signals (e.g. sentences or audio waveforms) are encoded as high-dimensional feature vectors by an artificial neural network; these features are then decoded by another neural network into predictions for a downstream task (e.g. text classification or machine translation). The encoder network can be trained without annotations via "self-supervision," typically to produce representations that

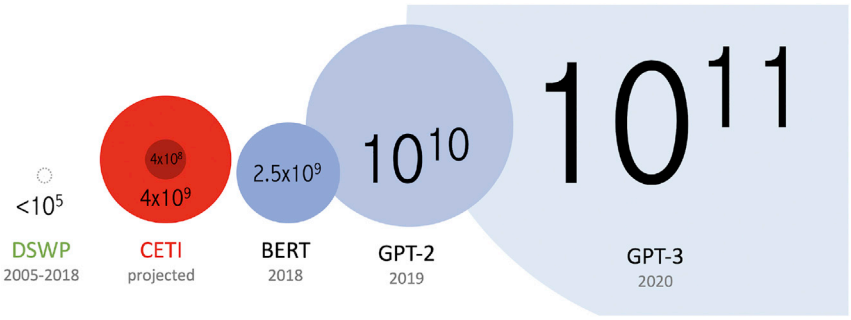

Figure 3. Comparative size of datasets used for training NLP models (represented by the circle area)
GPT-3 is only partially visible, while the DSWP dataset is a tiny dot on this plot (located at the center of the dashed circle). Shown in red is the estimated size of a new dataset planned to be collected in Dominica by Project CETI, an interdisciplinary initiative for cetacean communication interpretation. The estimate is based on the assumption of nearly continuous monitoring of 50–400 whales. The estimate assumes 75%–80% of their vocalizations constituting echolocation clicks, and 20%–25% being coda clicks. A typical Caribbean whale coda has five clicks and lasts 4 s (including a silence between two subsequent codas), yielding a rate of 1.25 clicks/sec. Overall, we estimate it would be possible to collect between 400M and 4B clicks per year as a longitudinal and continuous recording of bioacoustic signals as well as detailed behavior and environmental data.

make it possible to reconstruct parts of the input that have been hidden or corrupted. This apparently simple task requires a deep understanding of the structure of language and creates a rich language representation that can be used for a plethora of tasks, including automated grammar induction (Cao et al., 2020; Kim et al., 2020) and machine translation without parallel data (Lample et al., 2018).

However, a key characteristic of this self-supervision process is its reliance on massive collections of data: for example, the recent state-of-the-art Transformer models such as GPT-3 (Brown et al., 2020) was pre-trained on a large language corpus comprising over 10^{11} data points. While unsupervised structure discovery is also possible without self-supervised representation learning (Klein and Manning, 2004; Naseem et al., 2010), recent studies have also shown that unsupervised structure discovery can provide benefits (Harwath et al., 2020; Papadimitriou and Jurafsky, 2020).

It is difficult to make an exact analogy between tokens in human languages and whale vocalizations. And, for comparison, the Dominica Sperm Whale Project (DSWP) dataset contains less than 10^4 coda clicks (Figure 3) collected over a longitudinal study since 2005. It is thus apparent that one of the key challenges toward the analysis of sperm whale (and more broadly, animal) communications using modern deep learning techniques is the need for sizable datasets capturing a wide range of attributes. Secondly, human linguistic corpora are easier to deal with because they are typically *pre-analyzed* (i.e., already presented in the form of words or letters) and verification against ground truth is available, whereas in bioacoustic communication data, the relevant units must be inferred bottom-up with no ground truth available. Given this highly complex learning objective, we expect larger datasets will facilitate the discovery of meaning-carrying units.

RECORDING AND PROCESSING: BUILDING THE SPERM WHALE LONGITUDINAL DATASET

Data acquisition

Large-scale data collection over lengthy timespans (years of recordings and observation) requires the use of autonomous and semi-autonomous assets that continuously operate on, around, and above the whales (Figure 4). Multiple technologies available today can be utilized for purposes including localization of groups of sperm whales, time- and location-stamped audio recording, and collection of other data such as ocean conditions and video capturing of whales' behavior. Assets coming in contact with whales should be designed with non-invasive technology (Gruber and Wood, 2022) in order to minimize disturbance to animals, which in turn would provide more reliable data and also be more respectful to the study subjects. Finally, the location for data collection should ideally have a known large resident sperm whale population.

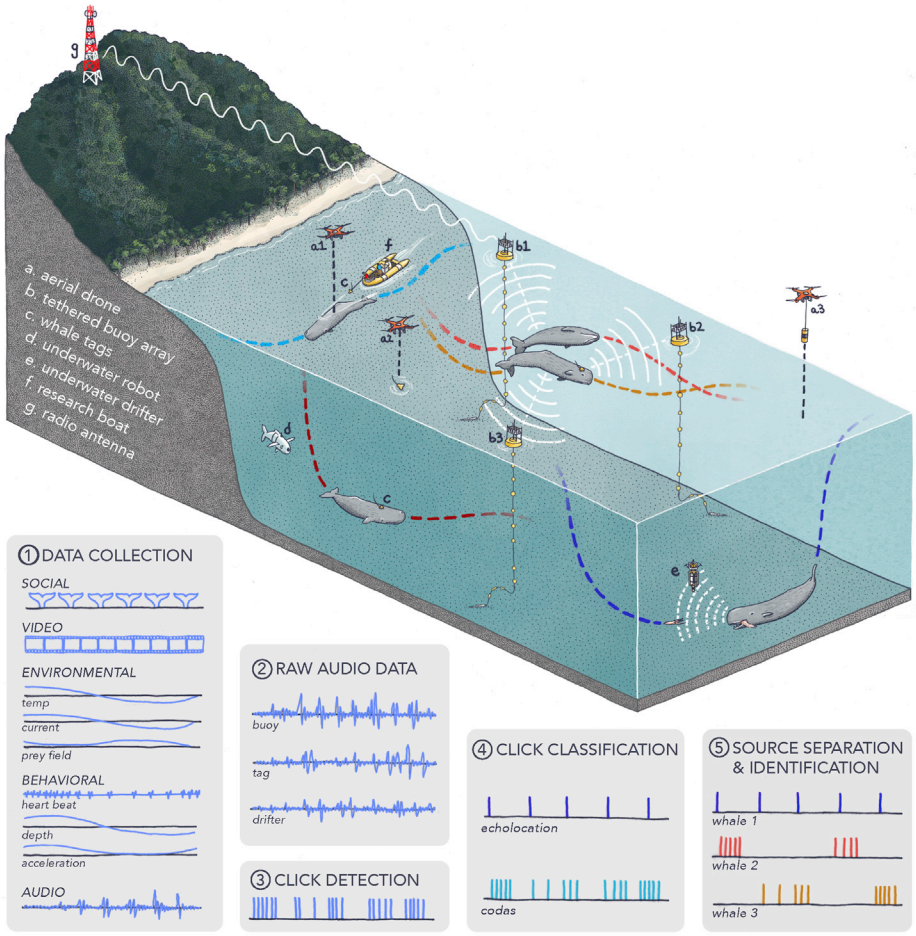

Figure 4. Schematic of whale bioacoustic data collection with multiple data sources by several classes of assets

These include tethered buoy arrays (b), which track the whales in a large area in real time by continuously transmitting their data to shore (g), floaters (e), and robotic fishes (d)Tags (c) attached to whales can possibly provide the most detailed bioacoustic and behavioral data. Aerial drones (a) can be used to assist tag deployment (a1), recovery (a2), and provide visual observation of the whales (a3). The collected multimodal data (1) have to be processed to reconstruct a social network of sperm whales. The raw acoustic data (2) have to be analyzed by ML algorithms to detect (3) and classify (4) clicks. Source separation and identification (5) algorithms would allow reconstructing multi-party conversations by attributing different clicks to the whales producing them. Illustration © 2021 Alex Boersma.

Tethered buoy arrays (Figure 4B) are a typical setup utilized for background recording of bioacoustic signals. Such installations usually comprise an array of sensors mounted at intervals of several hundred meters from the surface to the depth at which sperm whales are known to hunt, approximately 1200 m. The use of multiple sensors on each mooring and multiple moorings should allow the tethered arrays to localize the whales and track their movements. The advantage of such arrays is their reliability and capability to record signals continuously from a broad area in the ocean.

Tags (Figure 4C) or recording devices attached to whales have historically provided the most detailed insight into their daily activities and interactions (Johnson and Tyack, 2003). There are currently several designs of animal-borne recording devices that use suction to delicately attach to the whales and record not only the whale acoustics but also pressure, temperature, movement, and orientation. A critical current limitation of tags is onboard energy storage and memory as well as the effectiveness of their adhesion mechanisms. Bioinspired suction-based adhesion mechanisms inspired by carangiform fish (Gamel et al., 2019; Wang et al., 2017) and cephalopod tentacles hold the promise of achieving working times on the order of

several days and potentially to weeks. Fused with the sensor array data, the recordings from tags also allow the identification of whales in multi-party discourses and when associating behavior patterns with background recordings of the hydrophone/static sensor arrays.

Aquatic drones (Figures 4D and 4E): Free-swimming and passively floating aquatic drones allow obtaining audio and video recordings from multiple animals simultaneously to observe behaviors and communications within a group of whales near the surface. There is a wide spectrum of potential solutions from simple drifters to self-propelled robots capable of autonomous navigation, including numerous existing platforms that can be loosely categorized as active, submarine-like bodies or semi-passive "gliders". For self-propelled drones, small, short-range, bioinspired designs (Katzschmann et al., 2018; Marchese et al., 2014) hold the potential to operate in close proximity to a group of whales with minimal disruption.

Aerial drones (Figure 4A): Hybrid aerial/aquatic drones are capable of surveying areas to monitor the whale population, and providing "just-in-time" deployment of hydrophones and possibly also deploying and recovering tags. Current off-the-shelf drones have payloads in excess of several kilograms (in excess of our target tag mass) and flight times typically up to 30 min. This would allow an individual drone to cover an area with a radius of several kilometers for tag deployment and collection. Furthermore, amphibious drones with the ability to land on and take off from water can be used to directly carry and deploy recording devices to a site of interest.

Data processing

Given the large magnitude of data, a key step is to build appropriate data storage and processing infrastructure, including automated ML pipelines (maintainable and reusable across multiple data collecting devices) that will replace the annotation currently done largely by hand by marine biologists. ML-based methods are already being used for detection and classification among marine mammals (Gillespie et al., 2009; Shiu et al., 2020) and for sperm whale click detection and classification (Bermant et al., 2019; Ferrari et al., 2020; Glotin et al., 2018; Jiang et al., 2018); such methods are potentially scalable to large datasets containing years of recording that would otherwise be beyond reach with previous manual approaches.

By aggregating and synchronizing the bioacoustic, behavioral, and environmental signals from multiple assets (Figure 4), it is possible to localize the whales and continuously track them over time. The resulting dataset, a sort of "social network" of whales, will provide longitudinal information about the behavior and communications of individual whales (Farine and Whitehead, 2015; Sah et al., 2019; Sosa et al., 2021) and will be a crucial asset for subsequent machine learning.

DECODING AND ENCODING: BUILDING THE SPERM WHALE COMMUNICATION MODEL

In human languages, there has been substantial recent progress in automated processing and unsupervised discovery of linguistic structure, including acoustic representation learning (Chung et al., 2016; Kamper et al., 2014), text generation (Brown et al., 2020), induction of phrase structure grammars (Kim et al., 2019; Klein and Manning, 2004; Naseem et al., 2010), unsupervised translation (Artetxe et al., 2018; Lample et al., 2018), and grounding of language in perception and action (Lu et al., 2019; Shi et al., 2019), based on large-scale datasets. Similar tools could be applied to automatically identify structure in whale vocalizations.

Phonetics and phonology: Basic acoustic building blocks

One of the most striking features of human language is its discrete structure. While the sound production apparatus and the acoustic speech stream are fundamentally continuous (humans can modulate pitch, volume, tongue position, etc. continuously), human spoken languages partition this space into discrete units such as vowels, consonants, and tones (Eimas et al., 1971; Repp, 1984). Even though these discrete mental representations of sounds (*phonemes)* do not carry meaning, they form the building blocks from which larger meaning-carrying components are built. The distribution of phonemes in human languages is governed by a set of rules (phonotactic and phonological) that have also been identified, in a similar but simpler form, in vocalizations of non-human species, such as birds (Berwick et al., 2011).

Previous research has conjectured that sperm whale communication is also built from a set of discrete units. Codas—prototypical sequences of clicks with fixed relative inter-click interval structure—have been identified as such fundamental and discrete communicative units (Schulz et al., 2008; Weilgart and Whitehead,

1997). However, a plethora of questions remain. For example: are codas distinguished only by the absolute inter-click intervals, as suggested by previous studies? Do spectral features of coda clicks carry information? Does the frequency of individual clicks in codas carry meaning? What are the distributional restrictions (equivalents of phonotactic rules) governing codas and how can they be formalized (Antunes et al., 2011; Gero et al., 2016b)? Can we find equivalents of phonological computation in sperm whale vocalizations and what type of formal grammar best describes their vocalizations (Chomsky, 1956)? Answering these questions requires a combination of machine learning modeling as well as interpretable analytical approaches to the acoustic signal in order to understand how much information is lost when clicks are modeled as discretized units.

Identifying the fundamental units in whale vocalizations resembles spoken term discovery in human speech processing (Kamper et al., 2014), which has been addressed with a variety of unsupervised learning techniques (Baevski et al., 2020; Beguš, 2021; Chung et al., 2020; van Niekerk et al., 2020). These techniques use raw speech to automatically identify discrete clusters that represent repeated motifs—thereby finding structure inherent in the data via *compression*. Such techniques are already effective at automatically identifying words from raw audio of human languages (Baevski et al., 2020; Beguš, 2021; Chorowski et al., 2019; Chung et al., 2016; Eloff et al., 2019; van Niekerk et al., 2020). While learning representations in many of these models are not constrained to human speech, it is possible that dependencies in sperm whale communication diverge substantially from human speech, which means that both the models can learn misleading representations or our analysis of the outputs of the learned models will be influenced by anthropocentric biases. The risk of anthropocentrism in comparative animal behavior research is always present; a careful and un-biased study will be essential. One way to address this risk is to introduce context-specific multimodal data and model acoustic and behavioral data of sperm whales simultaneously (social and genetic relationships of signalers, behavioral budgets, foraging success, relative position in relation to conspecifics, velocity, orientation, pressure, water temperature, GPS information, weather etc.), which will provide the models information specific to whales.

Deep learning models for unsupervised discovery of meaningful units trained on human speech can readily be evaluated, inasmuch as independent identification of meaningful units in speech is almost always available. However, in the case of sperm whale vocalizations, validation is substantially more challenging and necessitates the use of behavioral data and playback experiments. Unsupervised learning is most effective when applied to large and diverse datasets (applications in human speech perform best with hundreds to thousands of hours of recordings (Chung and Glass, 2018)), highlighting the need for a large-scale bioacoustic data collection.

Morphology and syntax: Grammatical structure of communication

The capacity to construct complex words and sentences from simpler parts according to regular rules is one of the hallmarks of human language. While compositional codes appear in some animal communication systems (e.g. the waggle dance in honeybees composes independent distance and orientation factors (Glass, 2012), and Campbell's monkeys use affixation to alter alarm call meaning (Ouattara et al., 2009)), no known animal communication system appears to feature more complex structure-building operations like recursion, a central feature of almost all human languages. According to current knowledge, animal systems that have semantics (e.g. primate calls and gestures or bird calls) appear to have a simple syntax; on the other hand, systems that have a somewhat sophisticated syntax (e.g. birdsongs (Berwick et al., 2011)) are not associated with a compositional semantics.

In human languages, recent advances in NLP methods for *unsupervised grammar induction* (Kim et al., 2019; Klein and Manning, 2004; Naseem et al., 2010) have shown the possibility of accurately recovering dependency and phrase structure grammars from a collection of sentences. Applying such techniques to the discretized "basic unit" sequences of whale communications should allow for the generation of hypotheses about higher-level hierarchical structures across codas—the *syntax* of whale vocalization. As with the representation learning approaches for identifying basic units, large datasets are crucial for this effort: since any given sequence can be explained by many different candidate grammars, many sequences are necessary to adequately constrain the hypothesis space.

Semantics: Inferring meaning

Identifying short-term and long-term structure of vocalizations is a prerequisite to the key question: what do these vocalizations *mean*? The first step toward this goal is to identify the smallest meaning-carrying

units, analogous to *morphemes* in human languages. It is known that individual codas carry information about the individual, family, and clan identity (Antunes et al., 2011; Gero et al., 2016b; Oliveira et al., 2016), but the function of many codas, as well as their internal variability in structure and individual clicks, remains unexplained. It is imperative that the collected data used for machine learning captures this richer context of whale vocalizations, enabling the grounding of a wider set of morphemes and candidate meanings.

The grounding of minimal units ("morphemes"), together with identified hierarchical structures allows one to search for *interpretation rules*—associations of complex behaviors with long sequences of clicks via an explicit bottom-up process. A number of compositional semantic models in the NLP literature are capable of learning mappings between morpheme sequences and continuous groundings (Andreas et al., 2017; Socher et al., 2014). Currently existing whale bioacoustic datasets are likely too small for this purpose (see Figure 3), hence the need for acquiring a significantly larger and more detailed dataset. Finally, modeling composition should allow building a richer model of communicative intents, and ultimately to perform *interventional* studies in the form of playback experiments.

Discourse and social communication

Communication (whether human or non-human) occurs in a social context: speakers' reason about interlocutors' beliefs and intentions (Grice, 1975), explicitly signal the beginning and end of their conversational turns (Sacks et al., 1974), and adapt both the style and content of their messages to their audience (Giles et al., 1991). The complex, multi-party nature of sperm whale vocalization, and especially the presence of vocal learning and chorusing behaviors with no obvious analog in human communication (Patel, 2003; Schulz et al., 2008; Weilgart and Whitehead, 1993), suggests that this *discourse*-level structure is as important as the utterance-level structure for understanding whale communication.

Characterizing whales' *conversational protocols*, the rules that govern which individuals vocalize at what times, is key to understanding their discourse. Diverse communication protocols can be found across the animal kingdom—including uncoupled responding after a pause, chorusing in alternation, and chorusing synchronously—and each of these evolved protocols has been found to provide distinctive advantages for competitive or cooperative reproductive advantage, food advantage, and territorial defense (Ravignani et al., 2014). Variants of all these protocols have been observed in sperm whales (Schulz et al., 2008) and it is necessary to understand the roles that each of them plays vis-a-vis clan structure and group decision-making.

The understanding of conversational protocols is also a prerequisite to building *predictive models of conversations* (analogous to *language models* and *chatbots* for human-generated speech and text (Brown et al., 2020; Gao et al., 2019; Shannon, 1951)) capable of generating probable vocalizations given a conversation history, whale identities, and behavioral and environmental context. These models can be made controllable and capable of continuing vocalizations to express specific communicative intents (using inferred meanings for vocalizations in historical data) and will enable interactive playback studies.

Redundancy and fault tolerance of communication

Most forms of communication rely on the capacity to successfully transmit and receive a sequence of some basic units. In instances of imperfect acoustic channels with significant background noise, fault tolerance mechanisms are sometimes built into the communication system at different levels. In the animal kingdom, multiple fault tolerance mechanisms are known that exploit varying sensory modalities to backup communication signals (Johnstone, 1996), or adapt the communication units to noise conditions (LaZerte et al., 2016). Sperm whales, for example, have been shown to repeat vocalizations, including overlapping and matching codas (Schulz et al., 2008), a characteristic that might suggest redundancy mechanisms at the level of basic units and discourse. As studies venture into this area, it is important that such variations are distinguished from dialectal and individual variations, which can be detected using e.g. compression-based techniques (Oliveira et al., 2013).

Language acquisition

All human infants undergo similar stages during acquisition of language in their first years of life, regardless of the language in their environment. For example, the babbling period during which language-acquiring infants produce and repeat basic syllables (such as [da] or [ba]) or reduced handshapes and movements in

sign languages (Petitto and Marentette, 1991) is a well-documented developmental stage during the first 6–13 months (Fagan, 2009). Another well-documented concept in language acquisition is the critical period: if children are deprived of primary linguistic inputs in their first years, acquisition is not complete, often resulting in severe linguistic impairments (Friedmann and Rusou, 2015). The study of the developmental stages in language acquisition has yielded insights into how humans learn to discretize the acoustic speech stream into mental units, analyze meaning, and in turn produce language. In human language, for example, syllables that are produced first during language acquisition (e.g. [ma] or [ba]) are also most common in the world's languages, most stable, and easiest to produce. Similarly, morphological and syntactic constructions that are acquired first are the most basic (Crain and Thornton, 2012).

There are currently several known parallels in the developmental stages between human language and animal communication. Acquisition of birdsong in some species, for example, involves the presence of babbling as well as the critical period (Doupe and Kuhl, 1999). These parallels likely stem from common neural and genetic mechanisms behind human speech and animal vocalizations (Bolhuis et al., 2010; Musser et al., 2014). However, in cetacean research, existing data on the vocalizations of non-adult whales in their natural setting are limited. Continuous and longitudinal data acquisition capabilities are required to record vocalizations of calf-mother pairs and collect behavioral data on their interactions as calves mature. Such data will provide insights into the order of acquisition of coda types, leading to insights into the articulatory effort of the vocalization as well as identification of the most basic structural building blocks and their functions.

PLAYBACK-BASED VALIDATION

Playbacks are the experimental presentation of stimuli to animals, traditionally used to investigate their behavioral, cognitive, or psychophysiological responses (King, 2015). Playbacks in relation to animal communication can be categorized based on (i) the stimulus type (such as responses to conspecific or heterospecific signals (e.g. (Sayigh et al., 1999; Visser et al., 2016) or anthropogenic noise (e.g., sonar behavioral response studies, reviewed in the study by (Southall et al., 2016) and (ii) the collected data (such as response calls or behavior). While playback validation is a common technique used to study the vocalizations of terrestrial animals including birds (McGregor et al., 1992), primates (Fischer et al., 2013), and elephants (McComb et al., 2014; Stoeger and Baotic, 2016) that has proven successful in both grounding the functional use of calls as well as building understanding of the physiological and cognitive abilities of these animals in cetacean research, the vast majority of playback experiments have focused on the functional use of calls for social identity. It was shown this way, for example, that bottlenose dolphins use vocal labels to address one another (King et al., 2013).

For any vocal recognition system to function in this way, it must meet the following three criteria: first, there must be calls that vary enough and/or are sufficiently stereotyped to provide identity information for individuals or groups; second, listeners must be able to distinguish between these calls and hold a shared meaning/function for the calls; and third, listeners must then respond differently to those calls based on the identity of the signaler and their interaction history with them. There are hypotheses that off-axis portions of echolocation clicks can also carry information (Soldevilla et al., 2008).

The divide between playbacks *in situ* at sea and the captive experiments is partly a result of a separation in focus: captive studies have the capacity to examine the auditory capacity and cognitive responses of the animals, while wild studies can address the social and biological context of the response to conspecific calls. A good example among marine mammals is the two studies which when paired provided a holistic understanding of the semantics and function of signature whistles in bottlenose dolphin society. A captive playback study demonstrated that dolphins can learn and readily use vocal labels to address one another (King et al., 2013), while the field study demonstrated their use in this way at sea in social context among well-known individuals (Barber et al., 2001; Quick and Janik, 2012).

Another major separating factor is the logistical and technological limitations of performing playback studies at sea: studying animal communication in the wild within a natural social and behavioral context is significantly harder than in controlled settings in captivity. However, despite their complexity, wild playback experiments increase functional validity by avoiding the disturbance of species-typical social groups and daily behavioral routines (Cronin et al., 2017).

The inherent challenges of conducting playback experiments for the purpose of grounding hypotheses of any animal communication model fall under three general questions (see Deecke, 2006):

1) Do we know *what to playback*? Formalizing hypotheses requires a detailed understanding of both the signals being produced and the social/behavioral context in which they are used, and must be preceded by addressing core phonological, syntactical, and semantic questions using language models to better build appropriate playback stimuli to underlie grounding experiments within behavioral contexts.
2) Do stimuli *replicate biological signals*? Playback signals must adequately replicate the parameters of the natural signals themselves, avoid pseudo-replication with a sufficiently large sample, and reduce the logistical and perceptual limitations of conducting field playbacks from boats with researchers present. This requires developing playback technology based on autonomous interactive systems drifting at sea, which remove the vessel from the experiment, and have the capacity to listen and reply in context at biologically relevant speeds in order to approximate interactive playbacks (King, 2015).
3) Can we *recognize a response*? The ability to detect and identify behavioral responses to the playback stimuli requires a baseline understanding of the variation in behavior in the wild from observational studies. This is perhaps the biggest challenge as it requires both an understanding of what whales do, but also what we expect them to do in response to our playbacks.

While cetacean playbacks have similar interpretation challenges as terrestrial studies, they are logistically more challenging and mainly technologically limited. The purpose of playback experiments is 2-fold. First, and more typical, is the use of playbacks to ground semantic hypotheses and test purported syntax based on hypotheses generated from language models. The second use case, which can be viewed as an evolution of the first one, is more speculative but potentially offering an opportunity to make significant advancement in field-based, interactive playback among whales. There is currently rapid innovation of interactive playbacks in which researchers are able to more rapidly reply to animals' communication in the wild. This is particularly evident in bird song research (Dabelsteen and McGregor, 2020). Technological development of tools in this area in some ways mirrors advances with increasingly common NLP-based interactive voice assistants and chatbots, which are intended to listen, detect, and appropriately reply in context to their human users.

The ethical questions raised by playback studies are extensively covered in the study by (Cuthill, 1991) and deserve continued investigation and discourse. Playback experiments are, by design, active interaction with the animals under study. As such, playback paradigms and experimental design should endeavor to minimize potential impacts on the subjects and require focused natural observations before undertaking them. Playbacks have the potential to impact behavior over minutes or hours, as this is often the intent of the experiments to ground vocalizations in behavior, and mitigation measures should be in place and adverse behavioral responses are observed.

From the perspective of this study, cetaceans, especially whales, are undoubtedly disturbed due to the increased presence of anthropogenic underwater sounds, especially with increased global shipping trade. For example, Rolland et al. (2012) showed how the decrease in ship traffic following September 11, 2001 led to a significant reduction of stress-related fecal hormone metabolites in North Atlantic right whales. The aim of these studies is to not only better understand whale communication, but to also offer insights into how sound pollution impacts communication and behavior. Such information is critical when passing further legislation to better protect and conserve whales and other marine life.

Future steps

Recent advances in machine learning developed for the analysis of human-generated signals and broadly used in industry now make it possible to obtain unprecedented insights into the structure and meaning of non-human species communication. Such methods, when applied to purposely built datasets, are likely to bring a shift in perspective in deciphering animal communication in their natural settings. Achieving this ambitious goal requires an orchestrated effort and expertise from multiple disciplines across the scientific community. A prerequisite for this to happen is an open source and data sharing culture that has allowed the machine learning research community to flourish over the past decade. At present, there are promising

proposals to assemble a global library of underwater biological sounds collected by passive acoustic monitoring to catalog, study, and map the sounds made by underwater lifeforms worldwide (Parsons et al., 2022).

Previous large-scale and collaborative efforts have been successful in yielding substantial steps forward in the understanding of natural systems. Past collaborative projects (in particular, in genetics and astrophysics) turned out to be influential "not because they answer any single question but because they enable investigation of continuously arising new questions from the same data-rich sources" (Abbott et al., 2020), and their impact resulted from providing the technological foundations as well as findings and advancements along the journey.

Beyond advancing our understanding of natural communication systems, we see these efforts leading to tool sets that can be utilized in a diversity of fields. A large-scale, interdisciplinary, and integrated study of cetacean communication will also advance the design of underwater acoustic sensors, minimally invasive robotics, processing complex bioacoustic signals, and machine learning for language modeling. Collective advances in this area hold the potential to open new frontiers in interspecies communication and can lead to a deeper appreciation and understanding of the complexity and diversity of communication in the natural world.

CONSORTIA

The authors are the current scientific members of Project CETI collaboration, listed in alphabetical order, are Jacob Andreas, Gašper Beguš, Michael M. Bronstein, Roee Diamant, Denley Delaney, Shane Gero, Shafi Goldwasser, David F. Gruber, Sarah de Haas, Peter Malkin, Nikolay Pavlov, Roger Payne, Giovanni Petri, Daniela Rus, Pratyusha Sharma, Dan Tchernov, Pernille Tønnesen, Antonio Torralba, Daniel Vogt, and Robert J. Wood.

ACKNOWLEDGMENTS

We thank Jane Lipson, Pietro Liò, Philippe Schlenker, Emmanuel Chemla, and Camille Coye for helpful comments on the manuscript. This study was funded by grants from Dalio Philanthropies and Ocean X; Sea Grape Foundation; Rosamund Zander/Hansjorg Wyss, Chris Anderson/Jacqueline Novogratz through The Audacious Project: a collaborative funding initiative housed at TED; as well as support from National Geographic Society Grant (No. NGS-72337T-20) and Lyda Hill Philanthropies.

AUTHOR CONTRIBUTIONS

All authors wrote and reviewed the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Abbott, L.F., Bock, D.D., Callaway, E.M., Denk, W., Dulac, C., Fairhall, A.L., Fiete, I., Harris, K.M., Helmstaedter, M., Jain, V., et al. (2020). The mind of a mouse. Cell *182*, 1372–1376. https://doi.org/10.1016/j.cell.2020.08.010.

Ackers, S.H., and Slobodchikoff, C.N. (1999). Communication of stimulus size and shape in alarm calls of gunnison's Prairie dogs, Cynomys gunnisoni. Ethology *105*, 149–162. https://doi.org/10.1046/j.1439-0310.1999.00381.x.

Amano, M., Kourogi, A., Aoki, K., Yoshioka, M., and Mori, K. (2014). Differences in sperm whale codas between two waters off Japan: possible geographic separation of vocal clans. J. Mammal. *95*, 169–175. https://doi.org/10.1644/13-mamm-a-172.

Amorim, T.O.S., Rendell, L., Di Tullio, J., Secchi, E.R., Castro, F.R., and Andriolo, A. (2020). Coda repertoire and vocal clans of sperm whales in the western Atlantic Ocean. Deep Sea Res. Part I Oceanogr. Res. Pap. *160*, 103254. https://doi.org/10.1016/j.dsr.2020.103254.

Andreas, J., Dragan, A., and Klein, D. (2017). Translating neuralese. In 55th Annual Meeting of the Association for Computational Linguistics, pp. 232–242. https://doi.org/10.18653/v1/P17-1022.

Andreas, J., Rohrbach, M., Darrell, T., and Klein, D. (2016). Neural module networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 39–48.

Antunes, R., Schulz, T., Gero, S., Whitehead, H., Gordon, J., and Rendell, L. (2011). Individually distinctive acoustic features in sperm whale codas. Anim. Behav. *81*, 723–730. https://doi.org/10.1016/j.anbehav.2010.12.019.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018). Unsupervised neural machine translation. In International Conference on Learning Representations, pp. 1–12.

Baevski, A., Schneider, S., and Auli, M. (2020). vq-wav2vec: self-supervised learning of discrete speech representations. In International Conference on Learning Representations, pp. 1–12.

Baker, M.C. (2001). Bird song research: the past 100 years. Bird Behav. *14*, 3–50.

Balsby, T.J.S., and Bradbury, J.W. (2009). Vocal matching by orange-fronted conures (Aratinga canicularis). Behav. Process. *82*, 133–139. https://doi.org/10.1016/j.beproc.2009.05.005.

Barber, I., Arnott, S.A., Braithwaite, V.A., Andrew, J., and Huntingford, F.A. (2001). Indirect fitness consequences of mate choice in sticklebacks: offspring of brighter males grow slowly but resist parasitic infections. Proc. Biol. Sci. *268*, 71–76. https://doi.org/10.1098/rspb.2000.1331.

Beguš, G. (2021). CiwGAN and fiwGAN: encoding information in acoustic data to model lexical learning with Generative Adversarial Networks. Neural Netw. *139*, 305–325. https://doi.org/10.1016/j.neunet.2021.03.017.

Bermant, P.C., Bronstein, M.M., Wood, R.J., Gero, S., and Gruber, D.F. (2019). Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. Sci. Rep. *9*, 12588. https://doi.org/10.1038/s41598-019-48909-4.

Berwick, R.C., Okanoya, K., Beckers, G.J.L., and Bolhuis, J.J. (2011). Songs to syntax: the linguistics of birdsong. Trends Cogn. Sci. *15*, 113–121. https://doi.org/10.1016/j.tics.2011.01.002.

Bolhuis, J.J., Okanoya, K., and Scharff, C. (2010). Twitter evolution: converging mechanisms in birdsong and human speech. Nat. Rev. Neurosci. *11*, 747–759. https://doi.org/10.1038/nrn2931.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). language models are few-shot learners. Adv. Neural Inf. Process. Syst. *33*, 1–25.

Bruck, J.N. (2013). Decades-long social memory in bottlenose dolphins. Proc. Biol. Sci. *280*, 20131726. https://doi.org/10.1098/rspb.2013.1726.

Butti, C., Sherwood, C.C., Hakeem, A.Y., Allman, J.M., and Hof, P.R. (2009). Total number and volume of Von Economo neurons in the cerebral cortex of cetaceans. J. Comp. Neurol. *515*, 243–259. https://doi.org/10.1002/cne.22055.

Cantor, M., Gero, S., Whitehead, H., and Rendell, L. (2019). Sperm whale: the largest toothed creature on earth. In Ethology and Behavioral Ecology of Odontocetes, B. Würsig, ed. (Springer International Publishing), pp. 261–280.

Cantor, M., and Whitehead, H. (2015). How does social behavior differ among sperm whale clans? Mar. Mamm. Sci. *31*, 1275–1290. https://doi.org/10.1111/mms.12218.

Cao, S., Kitaev, N., and Klein, D. (2020). Unsupervised parsing via constituency tests. In EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp. 4798–4808. https://doi.org/10.18653/v1/2020.emnlp-main.389.

Chomsky, N. (1956). Three models for the description of language. IRE Trans. Inf. Theor. *2*, 113–124. https://doi.org/10.1109/tit.1956.1056813.

Chorowski, J., Weiss, R.J., Bengio, S., and van den Oord, A. (2019). Unsupervised speech representation learning using WaveNet autoencoders. In IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 2041–2053. https://doi.org/10.1109/TASLP.2019.2938863.

Chung, Y.A., and Glass, J. (2018). Speech2Vec: a sequence-to-sequence framework for learning word embeddings from speech. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. https://doi.org/10.21437/Interspeech.2018-2341.

Chung, Y.-A., Tang, H., and Glass, J. (2020). Vector-quantized autoregressive predictive coding. Preprint at arXiv. https://doi.org/10.48550/arXiv.2005.08392.

Chung, Y.-A., Wu, C.-C., Shen, C.-H., Lee, H.-Y., and Lee, L.-S. (2016). Audio Word2Vec: unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. Preprint at arXiv. https://doi.org/10.48550/arXiv.1603.00982.

Clarke, E., Reichard, U.H., and Zuberbühler, K. (2006). The syntax and meaning of wild gibbon songs. PLoS One *1*, e73. https://doi.org/10.1371/journal.pone.0000073.

Connor, R.C. (2000). Group living in whales and dolphins. In Cetacean Societies: Field Studies of Dolphins and Whales (University of Chicago Press), pp. 199–218.

Crain, S., and Thornton, R. (2012). Syntax acquisition. Wiley Interdiscip. Rev. Cogn. Sci. *3*, 185–203. https://doi.org/10.1002/wcs.1158.

Cronin, K.A., Jacobson, S.L., Bonnie, K.E., and Hopper, L.M. (2017). Studying primate cognition in a social setting to improve validity and welfare: a literature review highlighting successful approaches. PeerJ *5*, e3649. https://doi.org/10.7717/peerj.3649.

Cuthill, I. (1991). Field experiments in animal behaviour: methods and ethics. Anim. Behav. *42*, 1007–1014. https://doi.org/10.1016/S0003-3472(05)80153-8.

Dabelsteen, T., and McGregor, P.K. (2020). Dynamic acoustic communication and interactive playback. In Ecology and Evolution of Acoustic Communication in Birds (Cornell University Press), pp. 398–408.

Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., Tanburn, R., Battaglia, P., Blundell, C., Juhász, A., et al. (2021). Advancing mathematics by guiding human intuition with AI. Nature *600*, 70–74. https://doi.org/10.1038/s41586-021-04086-x.

Deecke, V.B. (2006). Studying marine mammal cognition in the wild: a review of four decades of playback experiments. Aquat. Mamm. *32*, 461–482. https://doi.org/10.1578/am.32.4.2006.461.

Doupe, A.J., and Kuhl, P.K. (1999). Birdsong and human speech: common themes and mechanisms. Annu. Rev. Neurosci. *22*, 567–631. https://doi.org/10.1146/annurev.neuro.22.1.567.

Eimas, P.D., Siqueland, E.R., Jusczyk, P., and Vigorito, J. (1971). Speech perception in infants. Science *171*, 303–306. https://doi.org/10.1126/science.171.3968.303.

Elias, D.O., Maddison, W.P., Peckmezian, C., Girard, M.B., and Mason, A.C. (2012). Orchestrating the score: complex multimodal courtship in the Habronattus coecatus group of Habronattus jumping spiders (Araneae: salticidae). Biol. J. Linn. Soc. Lond. *105*, 522–547. https://doi.org/10.1111/j.1095-8312.2011.01817.x.

Eloff, R., Nortje, A., Niekerk, B.V., Govender, A., Nortje, L., Pretorius, A., van Biljon, E., Westhuizen, E.V.D., van der Westhuizen, E., Staden, L.V., et al. (2019). Unsupervised acoustic unit discovery for speech synthesis using discrete latent-variable neural networks. In INTERSPEECH 2019, pp. 1–5. https://doi.org/10.21437/Interspeech.2019-1518.

Elsner, M., Goldwater, S., and Eisenstein, J. (2012). Bootstrapping a unified model of lexical and phonetic acquisition. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 184–193.

Fagan, M.K. (2009). Mean Length of Utterance before words and grammar: longitudinal trends and developmental implications of infant vocalizations. J. Child. Lang. *36*, 495–527. https://doi.org/10.1017/s0305000908009070.

Farine, D.R., and Whitehead, H. (2015). Constructing, conducting and interpreting animal social network analysis. J. Anim. Ecol. *84*, 1144–1163. https://doi.org/10.1111/1365-2656.12418.

Ferrari, M., Glotin, H., Marxer, R., and Asch, M. (2020). DOCC10: open access dataset of marine mammal transient studies and end-to-end CNN classification. In Proceedings of the International Joint Conference on Neural Networks, pp. 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207085.

Fischer, J., Noser, R., and Hammerschmidt, K. (2013). Bioacoustic field research: a primer to acoustic analyses and playback experiments with primates. Am. J. Primatol. *75*, 643–663. https://doi.org/10.1002/ajp.22153.

Fitch, W.T. (2005). The evolution of language: a comparative review. Biol. Philos. *20*, 193–203. https://doi.org/10.1007/s10539-005-5597-1.

Fögen, T. (2014). Animal communication. In The Oxford Handbook of Animals in Classical Thought and Life, G.L. Campbell, ed., pp. 1–18. https://doi.org/10.1093/oxfordhb/9780199589425.013.013.

Friedmann, N., and Rusou, D. (2015). Critical period for first language: the crucial role of language input during the first year of life. Curr. Opin. Neurobiol. *35*, 27–34. https://doi.org/10.1016/j.conb.2015.06.003.

Gamel, K.M., Garner, A.M., and Flammang, B.E. (2019). Bioinspired remora adhesive disc offers insight into evolution. Bioinspir. Biomim. *14*, 056014. https://doi.org/10.1088/1748-3190/ab3895.

Gao, J., Galley, M., and Li, L. (2019). Neural approaches to conversational AI. Found. Trends Inf. Retr. *13*, 127–298. https://doi.org/10.1561/1500000074.

Gero, S., Engelhaupt, D., Rendell, L., and Whitehead, H. (2009). Who cares? Between-group variation in alloparental caregiving in

sperm whales. Behav. Ecol. *20*, 838–843. https://doi.org/10.1093/beheco/arp068.

Gero, S., Gordon, J., and Whitehead, H. (2013). Calves as social hubs: dynamics of the social network within sperm whale units. Proc. Biol. Sci. *280*, 20131113. https://doi.org/10.1098/rspb.2013.1113.

Gero, S., Gordon, J., and Whitehead, H. (2015). Individualized social preferences and long-term social fidelity between social units of sperm whales. Anim. Behav. *102*, 15–23. https://doi.org/10.1016/j.anbehav.2015.01.008.

Gero, S., Bøttcher, A., Whitehead, H., and Madsen, P.T. (2016a). Socially segregated, sympatric sperm whale clans in the Atlantic Ocean. R. Soc. Open Sci. *3*, 160061. https://doi.org/10.1098/rsos.160061.

Gero, S., Whitehead, H., and Rendell, L. (2016b). Individual, unit and vocal clan level identity cues in sperm whale codas. R. Soc. Open Sci. 150372. https://doi.org/10.1098/rsos.150372.

Giles, H., Coupland, N., and Coupland, J. (1991). Accommodation theory: communication, context, and consequence. Contexts of Accommodation, pp. 1–68. https://doi.org/10.1017/CBO9780511663673.001.

Gillespie, D., Mellinger, D.K., Gordon, J., McLaren, D., Redmond, P., McHugh, R., Trinder, P., Deng, X.Y., and Thode, A. (2009). PAMGUARD: semiautomated, open source software for real-time acoustic detection and localisation of cetaceans. J. Acoust. Soc. Am. 2547. https://doi.org/10.1121/1.4808713.

Glass, J.R. (2012). Towards unsupervised speech processing. https://doi.org/10.1109/ISSPA.2012.6310546.

Glotin, H., Spong, P., Symonds, H., Roger, V., Balestriero, R., Ferrari, M., Poupard, M., Towers, J., Veirs, S., Marxer, R., et al. (2018). Deep learning for ethoacoustical mapping: application to a single Cachalot long term recording on joint observatories in Vancouver Island. J. Acoust. Soc. Am. *144*, 1776–1777. https://doi.org/10.1121/1.5067855.

Goldbogen, J.A., and Madsen, P.T. (2018). The evolution of foraging capacity and gigantism in cetaceans. J. Exp. Biol. *221*, jeb166033. https://doi.org/10.1242/jeb.166033.

Goyal, P., Duval, Q., Seessel, I., Caron, M., Misra, I., Sagun, L., Joulin, A., and Bojanowski, P. (2022). Vision models are more robust and fair when Pretrained on uncurated images without supervision. Preprint at arXiv. https://doi.org/10.48550/arXiv.2202.08360.

Grice, H.P. (1975). Logic and conversation. In Speech Acts (Brill), pp. 41–58.

Griesser, M., Wheatcroft, D., and Suzuki, T.N. (2018). From bird calls to human language: exploring the evolutionary drivers of compositional syntax. Curr. Opin. Behav. Sci. *21*, 6–12. https://doi.org/10.1016/j.cobeha.2017.11.002.

Gruber, D.F., and Wood, R.J. (2022). Advances and future outlooks in soft robotics for minimally invasive marine biology. Sci. Robot. *7*, eabm6807.

Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., and Glass, J. (2020). Jointly discovering visual objects and spoken words from raw sensory input. Int. J. Comput. Vis. *128*, 620–641. https://doi.org/10.1007/s11263-019-01205-0.

Haskell, J.P., Ritchie, M.E., and Olff, H. (2002). Fractal geometry predicts varying body size scaling relationships for mammal and bird home ranges. Nature *418*, 527–530. https://doi.org/10.1038/nature00840.

Hauser, M.D., Chomsky, N., and Fitch, W.T. (2002). Neuroscience: the faculty of language: what is it, who has it, and how did it evolve? Science *298*, 1569–1579. https://doi.org/10.1126/science.298.5598.1569.

Hebets, E.A., Vink, C.J., Sullivan-Beckers, L., and Rosenthal, M.F. (2013). The dominance of seismic signaling and selection for signal complexity in Schizocosa multimodal courtship displays. Behav. Ecol. Sociobiol. *67*, 1483–1498. https://doi.org/10.1007/s00265-013-1519-4.

Hof, P.R., Chanis, R., and Marino, L. (2005). Cortical complexity in cetacean brains. Anat. Rec. A Discov. Mol. Cell Evol. Biol. *287*, 1142–1152.

Huijser, L.A.E., Estrade, V., Webster, I., Mouysset, L., Cadinouche, A., and Dulau-Drouot, V. (2020). Vocal repertoires and insights into social structure of sperm whales (*Physeter macrocephalus*) in Mauritius, southwestern Indian Ocean. Mar. Mamm. Sci. *36*, 638–657. https://doi.org/10.1111/mms.12673.

Janik, V.M. (2014). Cetacean vocal learning and communication. Curr. Opin. Neurobiol. *28*, 60–65. https://doi.org/10.1016/j.conb.2014.06.010.

Janik, V.M., and Sayigh, L.S. (2013). Communication in bottlenose dolphins: 50 years of signature whistle research. J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol. *199*, 479–489. https://doi.org/10.1007/s00359-013-0817-7.

Janik, V.M., and Slater, P.J.B. (1998). Context-specific use suggests that bottlenose dolphin signature whistles are cohesion calls. Anim. Behav. *56*, 829–838. https://doi.org/10.1006/anbe.1998.0881.

Janik, V.M., and Slater, P.J.B. (1997). Vocal learning in mammals. Adv. Study Behav. *26*, 59–99. https://doi.org/10.1016/s0065-3454(08)60377-0.

Jiang, J.J., Bu, L.R., Wang, X.Q., Li, C.Y., Sun, Z.B., Yan, H., Hua, B., Duan, F.J., and Yang, J. (2018). Clicks classification of sperm whale and long-finned pilot whale based on continuous wavelet transform and artificial neural network. Appl. Acoust. *141*, 26–34. https://doi.org/10.1016/j.apacoust.2018.06.014.

Johnson, M.P., and Tyack, P.L. (2003). A digital acoustic recording tag for measuring the response of wild marine mammals to sound. IEEE J. Ocean. Eng. *28*, 3–12. https://doi.org/10.1109/joe.2002.808212.

Johnstone, R.A. (1996). Multiple displays in animal communication: 'backup signals' and 'multiple messages'. Philos. Trans. R. Soc. B Biol. Sci. *351*, 329–338. https://doi.org/10.1098/rstb.1996.0026.

Jones, C.B., and Van Cantfort, T.E. (2007). Multimodal communication by male mantled howler monkeys (Alouatta palliata) in sexual contexts: a descriptive analysis. Folia Primatol. *78*, 166–185. https://doi.org/10.1159/000099138.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kamper, H., Jansen, A., King, S., and Goldwater, S. (2014). Unsupervised lexical clustering of speech segments using fixed-dimensional acoustic embeddings. In 2014 IEEE Spoken Language Technology Workshop (SLT), pp. 100–105. https://doi.org/10.1109/SLT.2014.7078557.

Katzschmann, R.K., DelPreto, J., MacCurdy, R., and Rus, D. (2018). Exploration of underwater life with an acoustically controlled soft robotic fish. Sci. Robot. *3*, eaar3449. https://doi.org/10.1126/scirobotics.aar3449.

Kim, T., Choi, J., Edmiston, D., and Lee, S. (2020). Are pre-trained language models aware of phrases? Simple but strong baselines for grammar induction. In International Conference on Learning Representations (ICLR), pp. 1–22.

Kim, Y., Dyer, C., and Rush, A.M. (2019). Compound probabilistic context-free grammars for grammar induction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1906.10225.

King, S.L. (2015). You talkin' to me? Interactive playback is a powerful yet underused tool in animal communication research. Biol. Lett. *11*, 20150403. https://doi.org/10.1098/rsbl.2015.0403.

King, S.L., and Janik, V.M. (2013). Bottlenose dolphins can use learned vocal labels to address each other. Proc. Natl. Acad. Sci. U S A *110*, 13216–13221. https://doi.org/10.1073/pnas.1304459110.

King, S.L., Sayigh, L.S., Wells, R.S., Fellner, W., and Janik, V.M. (2013). Vocal copying of individually distinctive signature whistles in bottlenose dolphins. Proc. Biol. Sci. *280*, 20130053. https://doi.org/10.1098/rspb.2013.0053.

Klein, D., and Manning, C.D. (2004). Corpus-based induction of syntactic structure: models of dependency and constituency. In Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics, pp. 478–485.

Kulahci, I.G., Dornhaus, A., and Papaj, D.R. (2008). Multimodal signals enhance decision making in foraging bumble-bees. Proc. R. Soc. B Biol. Sci. *275*, 797–802. https://doi.org/10.1098/rspb.2007.1176.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Phrase-based & neural unsupervised machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 5039–5049.

LaZerte, S.E., Slabbekoorn, H., and Otter, K.A. (2016). Learning to cope: vocal adjustment to urban noise is correlated with prior experience in black-capped chickadees. Proc. Biol. Sci. *283*, 20161058. https://doi.org/10.1098/rspb.2016.1058.

Leavens, D.A. (2007). Animal cognition: multimodal tactics of orangutan communication. Curr. Biol. *17*, R762–R764. https://doi.org/10.1016/j.cub.2007.07.010.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444. https://doi.org/10.1038/nature14539.

Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1908.02265.

Marchese, A.D., Onal, C.D., and Rus, D. (2014). Autonomous soft robotic fish capable of escape maneuvers using fluidic elastomer actuators. Soft Robot. *1*, 75–87. https://doi.org/10.1089/soro.2013.0009.

Marcoux, M., Rendell, L., and Whitehead, H. (2007). Indications of fitness differences among vocal clans of sperm whales. Behav. Ecol. Sociobiol. *61*, 1093–1098. https://doi.org/10.1007/s00265-006-0342-6.

Marino, L. (2004). Cetacean brain evolution: multiplication generates complexity. Int. J. Comp. Psychol. *17*, 1–16. https://doi.org/10.1017/S0140525X00052961.

Marino, L. (1998). A comparison of encephalization between odontocete cetaceans and anthropoid primates. Brain Behav. Evol. *51*, 230–238. https://doi.org/10.1159/000006540.

Marino, L., Brakes, P., and Simmonds, M.P. (2011). Brain structure and intelligence in cetaceans. In Whales and Dolphins: Cognition, Culture, Conservation and Human Perceptions (EarthScan/Routledge), pp. 115–128.

McComb, K., Shannon, G., Sayialel, K.N., and Moss, C. (2014). Elephants can determine ethnicity, gender, and age from acoustic cues in human voices. Proc. Natl. Acad. Sci. U S A *111*, 5433–5438. https://doi.org/10.1073/pnas.1321543111.

McGregor, P.K., Catchpole, C.K., Dabelsteen, T., Bruce Falls, J., Fusani, L., Carl Gerhardt, H., Gilbert, F., Horn, A.G., Klump, G.M., Kroodsma, D.E., et al. (1992). Design of playback experiments: the Thornbridge Hall NATO ARW consensus. In Playback and Studies of Animal Communication (Springer), pp. 1–9.

Møhl, B., Wahlberg, M., Madsen, P.T., Heerfordt, A., and Lund, A. (2003). The monopulsed nature of sperm whale clicks. J. Acoust. Soc. Am. *114*, 1143–1154. https://doi.org/10.1121/1.1586258.

Musser, W.B., Bowles, A.E., Grebner, D.M., and Crance, J.L. (2014). Differences in acoustic features of vocalizations produced by killer whales cross-socialized with bottlenose dolphins. J. Acoust. Soc. Am. *136*, 1990–2002. https://doi.org/10.1121/1.4893906.

Naseem, T., Chen, H., Barzilay, R., and Johnson, M. (2010). Using universal linguistic knowledge to guide grammar induction. In Proceedings of EMNLP 2010: Conference on Empirical Methods in Natural Language Processing, pp. 1–11.

Oliveira, C., Wahlberg, M., Silva, M.A., Johnson, M., Antunes, R., Wisniewska, D.M., Fais, A., Gonçalves, J., and Madsen, P.T. (2016). Sperm whale codas may encode individuality as well as clan identity. J. Acoust. Soc. Am. *139*, 2860–2869. https://doi.org/10.1121/1.4949478.

Oliveira, W., Jr., Justino, E., and Oliveira, L.S. (2013). Comparing compression models for authorship attribution. Forensic Sci. Int. *228*, 100–104. https://doi.org/10.1016/j.forsciint.2013.02.025.

Ouattara, K., Lemasson, A., and Zuberbühler, K. (2009). Campbell's monkeys concatenate vocalizations into context-specific call sequences. Proc. Natl. Acad. Sci. U S A *106*, 22026–22031. https://doi.org/10.1073/pnas.0908118106.

Papadimitriou, I., and Jurafsky, D. (2020). Learning music helps you read: using transfer to study linguistic structure in language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 6829–6839. https://doi.org/10.18653/v1/2020.emnlp-main.554.

Parsons, M.J.G., Lin, T.H., Mooney, T.A., Erbe, C., Juanes, F., Lammers, M., Li, S., Linke, S., Looby, A., Nedelec, S.L., et al. (2022). Sounding the call for a global library of underwater biological sounds. Front. Ecol. Evol. *10*, 1–20. https://doi.org/10.3389/fevo.2022.810156.

Patel, A.D. (2003). Language, music, syntax and the brain. Nat. Neurosci. *6*, 674–681. https://doi.org/10.1038/nn1082.

Patterson, F.G. (1978). The gestures of a gorilla: language acquisition in another pongid. Brain Lang. *5*, 72–97. https://doi.org/10.1016/0093-934x(78)90008-1.

Pepperberg, I.M. (1990). Cognition in an African gray parrot (Psittacus erithacus): further evidence for comprehension of categories and labels. J. Comp. Psychol. *104*, 41–52. https://doi.org/10.1037/0735-7036.104.1.41.

Petitto, L.A., and Marentette, P.F. (1991). Babbling in the manual mode: evidence for the ontogeny of language. Science *251*, 1493–1496. https://doi.org/10.1126/science.2006424.

Poole, J.H., Tyack, P.L., Stoeger-Horwath, A.S., and Watwood, S. (2005). Elephants are capable of vocal learning. Nature *434*, 455–456. https://doi.org/10.1038/434455a.

Quick, N.J., and Janik, V.M. (2012). Bottlenose dolphins exchange signature whistles when meeting at sea. Proc. R. Soc. B Biol. Sci. *279*. https://doi.org/10.1098/rspb.2011.2537.

Ravignani, A., Bowling, D.L., and Fitch, W.T. (2014). Chorusing, synchrony, and the evolutionary functions of rhythm. Front. Psychol. *5*, 1118. https://doi.org/10.3389/fpsyg.2014.01118.

Rendell, L., Mesnick, S.L., Dalebout, M.L., Burtenshaw, J., and Whitehead, H. (2012). Can genetic differences explain vocal dialect variation in sperm whales, Physeter macrocephalus? Behav. Genet. *42*, 332–343. https://doi.org/10.1007/s10519-011-9513-y.

Rendell, L.E., and Whitehead, H. (2003). Vocal clans in sperm whales (Physeter macrocephalus). Proc. R. Soc. B Biol. Sci. https://doi.org/10.1098/rspb.2002.2239.

Repp, B.H. (1984). Categorical perception: issues, methods, findings. In Speech and Language, N.J. Lass, ed. (Elsevier), pp. 243–335.

Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., and Schiele, B. (2016). Grounding of Textual phrases in images by reconstruction. In Computer Vision – ECCV 2016 (Springer International Publishing), pp. 817–834.

Rolland, R.M., Parks, S.E., Hunt, K.E., Castellote, M., Corkeron, P.J., Nowacek, D.P., Wasser, S.K., and Kraus, S.D. (2012). Evidence that ship noise increases stress in right whales. Proc. R. Soc. B Biol. Sci. *279*, 2363–2368. https://doi.org/10.1098/rspb.2011.2429.

Sacks, H., Schegloff, E.A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. Language *50*, 7–55.

Sah, P., Méndez, J.D., and Bansal, S. (2019). A multi-species repository of social networks. Sci. Data *6*, 44. https://doi.org/10.1038/s41597-019-0056-z.

Savage-Rumbaugh, S., Rumbaugh, D.M., and McDonald, K. (1985). Language learning in two species of apes. Neurosci. Biobehav. Rev. *9*, 653–665. https://doi.org/10.1016/0149-7634(85)90012-0.

Sayigh, L.S., Tyack, P.L., Wells, R.S., Solow, A.R., Scott, M.D., and Irvine, A.B. (1999). Individual recognition in wild bottlenose dolphins: a field test using playback experiments. Anim. Behav. *57*, 41–50. https://doi.org/10.1006/anbe.1998.0961.

Schlenker, P., Chemla, E., Schel, A.M., Fuller, J., Gautier, J.-P., Kuhn, J., Veselinović, D., Arnold, K., Cäsar, C., Keenan, S., et al. (2016). Formal monkey linguistics. Theor. Linguist. *42*, 1–90. https://doi.org/10.1515/tl-2016-0001.

Schulz, T.M., Whitehead, H., Gero, S., and Rendell, L. (2011). Individual vocal production in a sperm whale (Physeter macrocephalus) social unit. Mar. Mamm. Sci. *27*, 149–166. https://doi.org/10.1111/j.1748-7692.2010.00399.x.

Schulz, T.M., Whitehead, H., Gero, S., and Rendell, L. (2008). Overlapping and matching of codas in vocal interactions between sperm whales: insights into communication function. Anim. Behav. *76*, 1977–1988. https://doi.org/10.1016/j.anbehav.2008.07.032.

Seim, I., Ma, S., Zhou, X., Gerashchenko, M.V., Lee, S.-G., Suydam, R., George, J.C., Bickham, J.W., and Gladyshev, V.N. (2014). The transcriptome of the bowhead whale Balaena mysticetus reveals adaptations of the longest-lived mammal. Aging *6*, 879–899. https://doi.org/10.18632/aging.100699.

Seyfarth, R.M., Cheney, D.L., and Marler, P. (1980). Vervet monkey alarm calls: semantic communication in a free-ranging primate. Anim.

Behav. *28*, 1070–1094. https://doi.org/10.1016/s0003-3472(80)80097-2.

Shannon, C.E. (1951). Prediction and entropy of printed English. Bell Syst. Tech. J. *30*, 50–64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x.

Shi, H., Mao, J., Gimpel, K., and Livescu, K. (2019). Visually grounded neural syntax acquisition. ACL Anthol. https://doi.org/10.18653/v1/P19-1180.

Shiu, Y., Palmer, K.J., Roch, M.A., Fleishman, E., Liu, X., Nosal, E.M., Helble, T., Cholewiak, D., Gillespie, D., and Klinck, H. (2020). Deep neural networks for automated detection of marine mammal species. Sci. Rep. *10*, 607–612. https://doi.org/10.1038/s41598-020-57549-y.

Slobodchikoff, C.N., Paseka, A., and Verdolin, J.L. (2009). Prairie dog alarm calls encode labels about predator colors. Anim. Cogn. *12*, 435–439. https://doi.org/10.1007/s10071-008-0203-y.

Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., and Ng, A.Y. (2014). Grounded compositional semantics for finding and describing images with sentences. Trans. Assoc. Comput. Linguist. *2*, 207–218. https://doi.org/10.1162/tacl_a_00177.

Soldevilla, M.S., Henderson, E.E., Campbell, G.S., Wiggins, S.M., Hildebrand, J.A., and Roch, M.A. (2008). Classification of Risso's and Pacific white-sided dolphins using spectral properties of echolocation clicks. J. Acoust. Soc. Am. *124*, 609–624. https://doi.org/10.1121/1.2932059.

Sosa, S., Jacoby, D.M.P., Lihoreau, M., and Sueur, C. (2021). Animal social networks: towards an integrative framework embedding social interactions, space and time. Methods Ecol. Evol. *12*, 4–9. https://doi.org/10.1111/2041-210x.13539.

Southall, B.L., Nowacek, D.P., Miller, P.J.O., and Tyack, P.L. (2016). Experimental field studies to measure behavioral responses of cetaceans to sonar. Endanger. Species Res. *31*, 293–315. https://doi.org/10.3354/esr00764.

Steele, J.H. (1985). A comparison of terrestrial and marine ecological systems. Nature *313*, 355–358. https://doi.org/10.1038/313355a0.

Stevick, P.T., Neves, M.C., Johansen, F., Engel, M.H., Allen, J., Marcondes, M.C.C., and Carlson, C. (2011). A quarter of a world away: female humpback whale moves 10,000 km between breeding areas. Biol. Lett. *7*, 299–302. https://doi.org/10.1098/rsbl.2010.0717.

Stoeger, A.S., and Baotic, A. (2016). Information content and acoustic structure of male African elephant social rumbles. Sci. Rep. *6*, 27585. https://doi.org/10.1038/srep27585.

Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., et al. (2020). A deep learning approach to antibiotic discovery. Cell *180*, 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021.

Suzuki, R., Buck, J.R., and Tyack, P.L. (2006). Information entropy of humpback whale songs. J. Acoust. Soc. Am. *119*, 1849–1866. https://doi.org/10.1121/1.2161827.

Tellex, S., Kollar, T., Dickerson, S., Walter, M.R., Banerjee, A.G., Teller, S., and Roy, N. (2011). Understanding natural language commands for robotic navigation and mobile manipulation. In Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI'11 (AAAI Press), pp. 1507–1514.

Tønnesen, P., Oliveira, C., Johnson, M., and Madsen, P.T. (2020). The long-range echo scene of the sperm whale biosonar. Biol. Lett. *16*, 20200134. https://doi.org/10.1098/rsbl.2020.0134.

Tyack, P. (1986). Population biology, social behavior and communication in whales and dolphins. Trends Ecol. Evol. *1*, 144–150. https://doi.org/10.1016/0169-5347(86)90042-x.

Tyack, P.L., and Sayigh, L.S. (1997). Vocal learning in cetaceans. In Social Influences on Vocal Development (Cambridge University Press), pp. 208–233.

van Niekerk, B., Nortje, L., and Kamper, H. (2020). Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge. In Interspeech 2020 (ISCA).

Visser, F., Curé, C., Kvadsheim, P.H., Lam, F.-P.A., Tyack, P.L., and Miller, P.J.O. (2016). Disturbance-specific social responses in long-finned pilot whales, Globicephala melas. Sci. Rep. *6*, 28641. https://doi.org/10.1038/srep28641.

Wang, Y., Yang, X., Chen, Y., Wainwright, D.K., Kenaley, C.P., Gong, Z., Liu, Z., Liu, H., Guan, J., Wang, T., et al. (2017). A biorobotic adhesive disc for underwater hitchhiking inspired by the remora suckerfish. Sci. Robot. *2*, eaan8072. https://doi.org/10.1126/scirobotics.aan8072.

Wanker, R., Sugama, Y., and Prinage, S. (2005). Vocal labelling of family members in spectacled parrotlets, Forpus conspicillatus. Anim. Behav. *70*, 111–118. https://doi.org/10.1016/j.anbehav.2004.09.022.

Watkins, W.A., and Schevill, W.E. (1977). Sperm whale codas. J. Acoust. Soc. Am. *62*, 1485–1490. https://doi.org/10.1121/1.381678.

Watwood, S.L., Miller, P.J.O., Johnson, M., Madsen, P.T., and Tyack, P.L. (2006). Deep-diving foraging behaviour of sperm whales (Physeter macrocephalus). J. Anim. Ecol. *75*, 814–825. https://doi.org/10.1111/j.1365-2656.2006.01101.x.

Weilgart, L., and Whitehead, H. (1997). Group-specific dialects and geographical variation in coda repertoire in South Pacific sperm whales. Behav. Ecol. Sociobiol. *40*, 277–285. https://doi.org/10.1007/s002650050343.

Weilgart, L., and Whitehead, H. (1993). Coda communication by sperm whales (Physeter macrocephalus) off the Galápagos Islands. Can. J. Zool. *71*, 744–752. https://doi.org/10.1139/z93-098.

Weilgart, L.S., Whitehead, H., and Payne, K. (1996). A colossal convergence - sperm whales and elephants share similar life histories and social structures, which include social females and roving males. Am. Sci. *84*, 278–287.

Whitehead, H. (2016). Consensus movements by groups of sperm whales. Mar. Mamm. Sci. *32*, 1402–1415. https://doi.org/10.1111/mms.12338.

Whitehead, H. (2003). Sperm Whales: Social Evolution in the Ocean (University of Chicago Press).

Whitehead, H., and Rendell, L. (2004). Movements, habitat use and feeding success of cultural clans of South Pacific sperm whales. J. Anim. Ecol. *73*, 190–196. https://doi.org/10.1111/j.1365-2656.2004.00798.x.

Worthington, L.V., and Schevill, W.E. (1957). Underwater sounds heard from sperm whales. Nature *180*, 291. https://doi.org/10.1038/180291a0.

Wosniack, M.E., Santos, M.C., Raposo, E.P., Viswanathan, G.M., and da Luz, M.G.E. (2017). The evolutionary origins of Lévy walk foraging. PLoS Comput. Biol. *13*, e1005774. https://doi.org/10.1371/journal.pcbi.1005774.

Zhong, M., Castellote, M., Dodhia, R., Lavista Ferres, J., Keogh, M., and Brewer, A. (2020). Beluga whale acoustic signal classification using deep learning neural network models. J. Acoust. Soc. Am. *147*, 1834–1841. https://doi.org/10.1121/10.0000921.

Zimmer, W.M.X., Tyack, P.L., Johnson, M.P., and Madsen, P.T. (2005). Three-dimensional beam pattern of regular sperm whale clicks confirms bent-horn hypothesis. J. Acoust. Soc. Am. *117*, 1473–1485. https://doi.org/10.1121/1.1828501.